2017

Maintaining a bitcoin address repository through focused web crawling

Caroline Wanjira Macharia
Faculty of Information Technology
Strathmore University

Maintaining a Bitcoin address repository through focused web crawling

Macharia Caroline Wanjira

Submitted in partial fulfillment of the requirements for the Degree of Masters in Information Systems Security at Strathmore University

Faculty of Information Technology

Strathmore University

Nairobi, Kenya

June, 2017

Declaration

I declare that this work has not been previously submitted and approved for the award of a degree by this or any other University. To the best of my knowledge and belief, the thesis contains no material previously published or written by another person except where due reference is made in the thesis itself.

..Macharia Caroline Wanjira……...

…….....

…….....

Approval

The thesis of Macharia Caroline Wanjira was reviewed and approved by the following:

Dr. Vladimír Veselý

Department of Information Systems

Brno University of Technology

Czech Republic

Dr. Joseph Orero

Dean, Faculty of Information Technology

Strathmore University

Dr. Ruth Kiraka

Associate Professor

Dean, School of Graduate Studies

Abstract

There has been an increase in the use of cryptocurrencies such as Bitcoin (BTC). Bitcoin allow for cross-border payments, for large and small items at little or no transaction fee. It is a groundbreaking technology that is not restricted by the current Anti-Money Laundering (AML) and Countering Financing of Terrorism (CFT). The decentralised nature of Bitcoin provides for anonymity through possible use of multiple identities that are not linked to personally identifiable information. Bitcoin offers a convenient solution for criminals who are restricted by the conventional money transfer systems. Bitcoin has been linked to drug dealers, gold bugs, fraudsters, terrorists, whistleblowers, pornographers, Internet freedom activists and unregulated gaming enterprises (Simser, 2015). Other cybercrimes that are facilitated by Bitcoin include ransomware.

This research identified a gap in acquiring digital evidence related to Bitcoin transactions. The data was obtained from case laws and secondary materials relating to Bitcoin forensics. We suggest a solution of maintaining a Bitcoin address repository through regular crawling of cryptocurrency sites. A web crawler that visits a list of user defined cryptocurrency sites was developed to solve the identified problem. The crawled sites include websites of interest to investigators. Results are stored in an XML file which can be exported to any database. The addresses and metadata collected from webpages can be used by investigators when building a case for Bitcoin related crimes. Transaction data about the collected addresses is available on the Bitcoin blockchain.

Keywords

Anonymity/ Privacy, Anti-Money Laundering (AML), Countering Financing of Terrorism (CFT) and terrorist financing, Law Enforcement Agencies, Bitcoin, Blockchain, Cryptocurrency, new payment methods, FATF, NFATs, Forensics, Bitcoin addresses

Table of Contents

List of Figures

Acknowledgement

I express my sincere and deep gratitude to my guide Dr. Vladimír Veselý, an assistant professor in Department of Information Systems at Brno University of Technology in Czech Republic, for the invaluable guidance and support. He provided me with resources and guidance for this dissertation.

Chapter 1 : Introduction to the Study

1.1.Introduction

The emergence of cryptocurrencies such as Bitcoin has led to new challenges in Anti-Money Laundering (AML) and Countering Financing of Terrorism (CFT). AML and CFT impose rules that make it hard for terrorist to make cross border payments with large sums of money. Bitcoin is a peer to peer electronic cash system (Nakamoto, 2016). An electronic coin or digital currency is money expressed as a string of bits sent as a message in a network that verifies the authenticity of the message via different mechanisms, such as proof-of-work (PoW) or proof-of-stake (PoS) (Tasca, 2015). Notable Bitcoin characteristics are: it is decentralised, easy to setup, requires minimal fee to transfer, it is anonymous, and it is non-repudiable. Transactions are publicly announced without linking them to anyone hence ensuring privacy. Privacy or anonymity and decentralisation pose a new challenge for law enforcement to track and prosecute criminals. Bitcoin technology is prone to problems experienced by other decentralised technologies such as Torrent File Distribution systems (Shusht & Dostov, 2014). Challenges with these technologies is they facilitate crime through anonymity. Furthermore, law enforcers are actively looking for ways to prevent decentralised networks from being abused by the criminals who distribute child porn, trade secrets or other illegal materials.

This dissertation focuses on the greatest challenge in Bitcoin which is its preferential use by terrorists and other criminals to conduct untraceable transactions. Existing Anti-Money Laundering and Counter Financing Terrorism laws act as a hindrance for crime funding in the traditional finance system. Cryptocurrencies such as Bitcoin, alternative coins such as Ethereum, Ripple offer possibilities to criminals due to their attribute of anonymity and decentralisation. Bitcoin is the most dominant cryptocurrency (Simser, 2015).

Forensic processes are used in the digital investigation process. The concept of anonymity is core to Bitcoin and other cryptocurrencies. Bitcoin and other cryptocurrencies offer anonymity by offering multiple identities that are not linked with personally identifiable information. The Bitcoin technology allows for all transactions to be publicly announced. Privacy is maintained by breaking the flow of information in another place: by keeping

public keys anonymous. The public can see that someone is sending an amount to someone else, but without information linking the transaction to anyone. This is similar to the level of information released by stock exchanges, where the time and size of individual trades “("tape")” is made public, but without telling who the parties were.

With the existing AML/CFT in traditional financial systems, it is becoming increasingly difficult for a terrorist to transfer money across jurisdictional borders (Shusht & Dostov, 2014). Digital currencies offer a convenient solution for cyber criminals. Enforcement agencies have raised alarms about Bitcoin as the cryptocurrency of choice on deep websites like Silk Road which is a black market for drugs (Cottle, 2016).

1.2. Problem Statement

Bitcoin identities and transactions are meant to be anonymous in nature which makes it prone misuse. There is a danger of using it for illegal goods trader, ransomware or terrorist funding. Investigators are faced with a challenge of gathering evidence to prosecute criminals.

1.3. General Objective

To develop a forensic tool that will regularly visit top cryptocurrency websites, parse, gather and store addresses with their website occurrences in the database. The stored data should provide as much information as possible to the investigators to assist them build a case.

1.4. Objectives

1. To understand the challenge posed by Bitcoin’s anonymity in tackling money laundering, terrorism financing and other crimes such as ransomware.
2. To provide research on existing solutions handling money laundering and terrorism funding using digital currencies; identify a gap in the forensic investigation of Bitcoin-related crimes.
3. To design, develop and test an application that extracts Bitcoin’s information and can be queried by network administrators or LEA representatives.

4. To validate the developed tool to ensure it tackles the challenge of evidence acquisition when investigating Bitcoin related crime.

1.5. Research Questions

1. What challenge is posed by decentralised payment system such as Bitcoin?
2. Which solutions are available to tackle Bitcoin forensics and which gap can be addressed?
3. What features will be integrated into the developed application?
4. Does the developed application address the challenge of evidence acquisition needed for Bitcoin forensic investigations?

1.6. Scope and Limitations

The scope of the system to be developed will work on the current version of Bitcoin v1 protocol. Even though there exist many cryptocurrencies, this research will only focus on Bitcoin since it is currently the most widely used cryptocurrency.

1.7. Justification

Cryptocurrencies offer a groundbreaking solution in the finance industry. Despite its numerous benefits, they can be used to the detriment of society. Child pornography, terrorism, and illegal goods trade are some of the crimes that utilise the dark web to maintain anonymity. Other cybercrimes that are on the rise include ransomware facilitated by digital payments.

Evidence suggests that Bitcoin has been utilised in some successful terror attacks. The first was in two bombing attacks conducted by an ISIS-inspired lone wolf on a shopping mall in Jakarta, Indonesia in 2015 and the second was in the November 2015 coordinated terror attacks in France (Cottle, 2016). Ghost Security Group claims that they analysed an enormous amount digital data in the aftermath of the French attacks in an attempt to track the movements of those responsible. They report that their investigations of the attackers led them to a Bitcoin wallet containing $3 million (Milad & Irwin, 2016).

Chapter 2: Literature Review

2.1. Cryptocurrencies

2.1.1. Overview

A cryptocurrency or a digital currency is a digital asset that works as a medium of exchange using cryptography to secure transactions. Cryptocurrencies are monetary value expressed as a string of bits sent as a message in a network that confirms the originality of the message via different mechanisms, such as proof-of-work (PoW) or proof-of-stake (PoS). Bitcoin is the first decentralised cryptocurrency and it was released in 2009 by Satoshi Nakamoto (Milad & Irwin, 2016). There total number of cryptocurrencies is around 637. The current Bitcoin market capitalisation is $20,846,606,998 and Bitcoin has an 85% dominance. Top five digital currencies are Bitcoin, Ethereum, Ripple, Litecoin, Dash, Monero as ranked by market capitalisation (Crypto-Currency Market Capitalizations, 2016).

Bitcoin transactions are recorded in a blockchain. A blockchain is a distributed database that maintains a continuously-growing list of data records hardened against tampering and revision (Blockchain, 2016). It consists of data structure blocks which hold exclusively data in initial blockchain implementations, and both data and programs in some of the more recent implementations with each block holding batches of individual transactions and the results of any blockchain executables. A block is a file that stores transaction data (Wiki: Block, 2016). New transactions are continuously added to new blocks. Each block contains a timestamp and information linking it to a previous block (Blockchain, 2016). Blockchain has resulted in a reliable way to handle interactions requiring trust, proof, and contract enforceability. Technological disruptions are enabling a shift towards what is known as the economy of Collaborative Commons (CC): a digital space where providers and users share goods and services at close to zero marginal cost. These technologies are referred as blockchain technologies (Cottle, 2016).

Advancement in cryptography, network connectivity and computing power has led to development of infrastructures that allow for the creation of trust-less disintermediated and decentralised markets. Potential benefits of blockchain will be realised in

intermediation, clearing and settlement, recording systems, rating or voting systems, transaction traceability schemes and many others. Bitcoin is the first digital decentralised cryptocurrency powered by blockchain. Transparency means transactions can be processed without the need for third-party verification.

We are transitioning from an era of e-finance into peer to peer (P2P) or decentralised finance. With decentralised finance, financial services are directly provisioned by end users to end users using computer-based and network-based information and communication technologies. Decentralised finance is based on information communication technologies, cryptography, open source computing methods, time-stamped ledgers, and peer-to-peer distributed networks that allow end-users direct anonymous, disintermediated and secure access to assets, payments, and financial services. Benefits of decentralised finance are that they will reduce the costly intermediation layers, foster financial inclusion and reduce cash dependence. Like other decentralised technologies such as Torrent File Distribution system, it also poses new challenges for legal systems, issues related to safety and soundness of regulation and competition policy, consumer and investor protection and global public policy. Emerging challenges resulting from decentralised finance include: emergence of new crimes such as fraud and cybercrimes; uncontrolled herding behaviour and excessive firm concentration, thereby creating market distortions with potentially disastrous repercussions on the real economy; bubble dynamics and preferential use by terrorists and criminals to conduct untraceable transactions (Shusht & Dostov, 2014).

The legality of cryptocurrencies varies from country to country. Russia has accepted cryptocurrencies as legal, the limitation is that it is illegal to purchase goods with any currency other than the Russian ruble. The United States Internal Revenue Service ruled that Bitcoin will be treated as property for tax purposes as opposed to currency (Smith, 2014). This implies that it is subject to capital gain tax. The latest pronouncements by Financial Crimes Enforcement Network (FinCEN) analysed in connection with Superintendent Law that US federal and state governments are not trying to stifle or control virtual currencies. On the contrary, these efforts demonstrate that US

governmental entities recognise the long-term value of virtual currencies and are trying to create a regulatory regime to foster growth and development, and an atmosphere where institutional and retail investors are protected (Xethalis, 2015).

2.1.2. Bitcoin

Bitcoin was designed on an open source cryptographic protocol platform which enables each transaction between the users to be "computationally impractical" to reverse and would protect sellers from fraud (Nakamoto, 2008).

A Bitcoin network consists of clients (who can be users and/ or, miners) that generate transactions. Miners approve transactions on a network by using special software by solving complex math problems (What is Bitcoin Mining?, 2017). A client manages wallet and contained addresses. Client software provides services of the Bitcoin network. Transaction tracks how Bitcoin are passed between users. A Bitcoin uses a public ledger known as blockchain. Every node in the network collectively composes a "decentralised" bank that is the blockchain. Each node downloads new blocks, keeps track of who has which Bitcoin, verifies that transactions are legitimate and registers in the ledger the passage of messages between users (Nakamoto, 2016).

Bitcoin transactions are sent from and to electronic Bitcoin wallets, and are digitally signed for security (CoinDesk, 2015). Transactions have inputs; records that have references from previous transactions and, output: this is the record which determines the new owner of the transferred Bitcoin. Bitcoin address is used to send Bitcoin to the recipient (CoinDesk, 2015).

1. Each input must have a cryptographic digital signature that unlocks the funds from the prior transaction. Only a person with the appropriate private key can create a satisfactory signature (Bitcoinwiki: Protocol documentation, 2016).
2. Output: determines address or addresses of the recipient or recipients (CoinDesk, 2015).

In a transaction, the sum of all inputs must be equal to or greater than the sum of all outputs. If the inputs exceed the outputs, the difference is considered a transaction fee, and is redeemable by whoever first includes the transaction into the block chain. For A to

send Bitcoin to B, he needs B's address; Bitcoin address is automatically generated by the wallet software.

Mining is the process of verifying new transactions to prevent double spending (Bitcoin Miner, 2015). It involves adding verified transaction records to Bitcoin's public ledger. Mining is programmed to be resource-intensive and tasking so that the number of blocks found each day by miners remains steady. Individual blocks must contain a proof of work to be considered valid. This proof of work is verified by other Bitcoin nodes each time they receive a block. Bitcoin uses the hashcash proof-of-work function. Hashcash is a proof-of-work algorithm, which has been used as a denial-of-service countermeasure technique in several systems. The server generates a string called a stamp that contains encoded information about the client. A sender computes the hashcash stamp while the recipient or anyone in the public can verify received hashcash stamps efficiently (Hashcash, 2016). The mining reward started at 50 Bitcoin and is expected to halve in every 210,000 blocks. There is a cap of 21 million Bitcoin that will ever be mined (Wiki: Block, 2016). Mining should increasingly get difficult as we get closer to the cap.

2.1.3. Traceability and Accountability

In the traditional banking model, privacy is achieved by limiting access to information to the parties involved and the trusted third party. With Bitcoin, transactions are publicly announced but privacy is maintained by breaking the flow of information in another place: by keeping public keys anonymous. The public can see that someone is sending an amount to someone else, but without information linking the transaction to anyone. This is like the level of information released by stock exchanges, where the time and size of individual trades, the "tape", is made public, but without telling who the parties were (Nakamoto, 2016)

Bitcoin transactions are made public while identities are private. The concept of anonymity is controversial in Bitcoin. According to Doran (2015), they are pseudonymous. Pseudonymity is the concept of disguised identity. He argues that the Bitcoin network is one of the most transparent payment system. Bitcoin do not just appear; each transaction is public knowledge and is visible to anyone using the Bitcoin network.

The absence of personally identifiable information (PII) which links sellers and buyers to actual transactions achieves anonymity (Milad, 2016). For extra invisibility, a new key pair should be used for each transaction to keep them from being linked to a common owner. A new keypair is generated for each receiving address. A Bitcoin address is a 160-bit hash of the public portion of a public/private Elliptic Curve Digital Signature Algorithm (ECDSA) keypair. With newer wallets, addresses are generated deterministically (Technical background of version 1 Bitcoin addresses, 2016). Bitcoin addresses contain a built-in check code, which prevents a user from sending to the wrong address. In a multi-input transaction, linking is unavoidable with multi-input transactions. If the owner of a key is revealed, the risk is that linking could reveal other transactions (Protect your privacy, 2016).

Some researchers argue that Bitcoin is wrongly perceived as an anonymous payment network (Protect your privacy, 2016). Bitcoin is arguably the most transparent payment network in the world. At the same time, Bitcoin can provide acceptable levels of privacy when used correctly. An individual ought to adopt good practices to protect their privacy. Good practices include using new addresses for new payments, hiding IP addresses with tools like Tor. By 2012, Federal Bureau of Investigation (FBI) was expressing difficulty in tracking the identity of suspicious anonymous Bitcoin users (Milad & Irwin, 2016).

2.1.4. Bitcoin Protocol

Bitcoin uses three different protocols for its functionality. Bitcoin v1 protocol runs on TCP port number 8333. Bitcoin v1 is employed for P2P communication between peers (Pluskal, et al., 2015). The process of peer discovery is provided twice in a network. When a client is started for the first time, programs do not know the IP addresses of any active full nodes. To discover some IP addresses, they query one or more well-known DNS names (called DNS seeds). The response to the lookup should include one or more DNS A records with the IP addresses of full nodes that may accept new incoming connections (Bitcoin Developer Guide, 2016).

Once a client is connected to a Bitcoin, the client may request for more peers from its neighbors to increase its connectivity graph. Protocol messages used in the protocol

include: version and verack (useful for connection initiation), address (to detect a communication graph and provide information of known nodes), and ping-pong (a keep-alive mechanism). Messages such inv, tx, and block transmit valuable information about processed transactions. The second group of protocol is used for work distribution among miners. It consists of Getwork, Getworktemplate, Stratum. Last group involves remote procedure call (RPC) messages that are employed for remote control of various Bitcoin related services (e.g., remote wallets controlled by a smartphone, on-line trading on Bitcoin exchanges) (Pluskal, et al., 2015).

2.1.5. Bitcoin Artefacts

An artefact is something observed in a scientific investigation or experiment that is not naturally present but occurs because of the preparative or investigative procedure (Oxford Dictionary, 2017). Bitcoin transactions are identifiable within the user's Bitcoin "wallet" by a specific address characterised by a 34-alphanumeric character string. The address is unique to a user's wallet and denotes specific transactions between various users (Shaw, 2013)

All Bitcoin transactions information is transparent. Blockchain contains all the information about Bitcoin money supply. An individual can verify their transactions or signatures associated with transaction on blockchain websites and conduct a search based on block number, address, block hash, the amount, the type, transaction hash or public key and balance available to the Bitcoin address (Block Explorer, 2016).

For instance, a Bitcoin Qt wallet consists of three folders; 'block", 'database", and "chainstate". Additionally, it has the following files: 'lock", "db'log", "debug.log", "peers.dat", "wallet.dat". Each file serves a specific purpose, "wallet.dat file stores keys, transactions, metadata and options relating to Bitcoin. On the client device, all the information pertaining to a user's Bitcoin account is stored in the "wallet.dat file". Internet Evidence Finder from Magnet can analyse a folder with Bitcoin artefacts or the whole image from a suspect's machine (Doran, 2015). It recovers "wallet.dat" file and present addresses from a Bitcoin wallet, as well as queries to the Bitcoin network from log files created by the Bitcoin client software in a user-friendly format (Doran, 2015). Internet

search history plays an important role too. Evidence that shows a client visiting Bitcoin related sites can help an investigator in building a case.

From a network perspective, Bitcoin detection involves identifying the TCP port number 8333. Protocol messages such as version, verack show that a connection is being initiated. Ping-pong is a keep alive mechanism in the Bitcoin network to check on peers. Ping message is sent primarily to confirm that the TCP/IP connection is still valid while pong message is sent in response to a ping message (Bitcoin wiki, 2016). Detection of ping pong shows the existence of Bitcoin nodes in the network. Detection of Getwork, Getworktemplate, Stratum protocols implies an existence of Bitcoin miner in the local network. There are many Remote procedural call messages used for Bitcoin related services. These messages indicate Bitcoin related services are running on the network (Pluskal, et al., 2015).

2.1.6. Decentralised Identity Management

The concept of decentralised identity management does away with the need for a central authority. A user can generate a new identity at any time. If you want to maintain anonymity, you can generate a new identity for some time and later discard it. These identities in Bitcoin are known as addresses, normally a hash of public key (Arvind, Edward, Andrew, & Steven, 2015).

The Bitcoin address is necessary when conducting transactions; which by themselves do not reflect identifying information (Brill & Keene, 2014). Addresses are random numbers which consist of a string of digits and characters that can be shared with anyone and used to send Bitcoins to. Addresses are produced from public keys by using one-way cryptographic hash functions. For each address one can do a cash flow analysis and its balance can be observed at each point by verifying historical transaction that involves specified address as input or output. Given an address A one can verify from which other addresses B, C, etc. it got the coins from and to which other addresses D, E, etc. it sent the coins to (Tasca, 2015). The transfer of value is permanently recorded on the blockchain (Brill & Keene, 2014).

Ownership of Bitcoins is established through the possession of a value k that is automatically generated (also offline and independently from the Bitcoin protocol) and stored in a file called a wallet via software called a Bitcoin client. A must keep k secret always, as revealing it to third parties is equivalent to giving them control over the Bitcoins secured by k. Any accidental loss of k must also be prevented, otherwise funds linked and secured by it are lost as well (Bitcoin Developer Guide, 2017).

The Bitcoin address is an identifier of 26-35 alphanumeric characters, beginning with the number 1 or 3, that represents a possible destination for a Bitcoin payment. Addresses can be generated at no cost by any user of Bitcoin. The two-address format are: P2PKH which starts with a 1 e.g. 1BvBMSEYstWetqTFn5Au4m4GFg7xJaNVN2 and P2SH type starts with a 3 3J98t1WpEZ73CNmQviecrnyiWrnqRhWNLy. They consist of random digits and uppercase and lowercase letters, with the exception that the uppercase letter "O", uppercase letter "I", lowercase letter "l", and the number "0" are never used to prevent visual ambiguity. Several characters in the address are used as checksum so that typographical errors can be automatically found and rejected (Bitcoinwiki : Address, 2017).

Unlike e-mail addresses, a Bitcoin address can be used once and people have many different Bitcoin addresses. A unique address should be used for each transaction. Addresses can be created offline. Addresses are case sensitive and exact and best practice is to copy paste the address when using them. Use of different addresses reduces chances of traceability and makes the forensic process complicated. Most popular addresses can be determined by the number of outputs e.g. SatoshiDice, LuckyBit and Deepbit (Blockchain.Info - Popular Addresses, 2017).

There is a false impression of total anonymity and privacy in Bitcoins. Even if the address is not linked to an individual it displays a pattern of behaviour. On the blockchain, people see the statements and thus know that whoever owns this identity has done a certain series of actions. When investigating, the address can guide investigators determine the activities and correlate it with a crime. By connecting the dots, a shrewd investigator can infer things about your real-world identity (Arvind, Edward, Andrew, & Steven, 2015). Investigators

also collaborate with Bitcoin exchange markets to obtain more information on Bitcoin transactions (Vesely, 2016).

2.1.7. Bitcoin's Preferential Use by Criminals

Bitcoins and other cryptocurrencies are an attractive option for cyber criminals. Features such as transaction anonymity and irreversibility payments make them ideal for use by money launderers, drug dealers and funding terrorists. A criminal aim is to leave the least footprint possible when conducting transactions.

Anonymity is key for criminals, the preference is a payment system that does not require identification. Online anonymity has proved beneficial not only for criminals but also governments and investigative bodies. It is beneficial when one is protecting their identity. With Bitcoin, transactions are publicly announced without linking them to an individual (Arvind, Edward, Andrew, & Steven, 2015).

Global reach has been made possible through Bitcoin. This allows for global participation from all countries including marginalised nations. Identities are not needed to complete transactions. Speed is important when conducting suspicious transactions to ensure that the transaction is not intercepted or blocked. For non-repudiation, no additional verification or validation should be necessary to execute any transaction. The person sending the money should not be able to 'unsend' it or reverse the transfer (Simser, 2015).

Low cost to use, though it is not a big concern for criminal, it is desirable if the system operates with minimal overhead allowing large and small transactions to occur without eating up the value of small transactions in fees. Lack of intermediaries lowers the cost of transaction. Relative ease of use is dependent on the graphical user interface; it should be easy for non-technical people to use. Preferably it should have a computer interface that makes setting up a transaction fast and easy and should be able to be used on a computer, tablet or smartphone that is connected to the Internet (What is Bitcoin Mining?, 2017).

Difficult for authorities to track transactions, the Internal Revenue Service (IRS) and FBI have expressed challenges with investigating Bitcoin related crimes. Sites such as Silk

Road that was dealing with illegal good trade took time and effort to be cracked down (Silk Road (marketplace), 2017).

Potential upgrades to security and anonymity, Bitcoin core developers are consistently improving on the protocol. Using additional layers of anonymity (through anonymising networks that currently exist or could be created) would make the job of law enforcement and anti-terrorism agencies much harder. While some jurisdictions have taken measures to regulate Bitcoin use most are yet to set up regulations. This means that a criminal can take advantage of the lesser strict countries because the cyber space is borderless (Nelson, Phillips, & Steuart, 2014).

2.2. Forensic Guidelines for Digital Investigations

A case is successfully prosecuted based on the admissibility of evidence recovered at the scene and the evidence extracted from the analysis of each piece of digital media. Effective digital investigation follows specific processes. Processes include: analysis, identification, preparation, approach strategy, preservation, collection, examination and analysis. For network, forensic analysis, data is acquired from sensors used to collect traffic data. Sensors used must be secure, fault tolerant and have limited access. A well defined procedure using reliable tools, hardware and software, must be in place to gather maximum evidence causing minimum impact to the victim (Nelson, Phillips, & Steuart, 2014).

The integrity of data logged and network events recorded must be ensured. Evidence presented in a court of law must be admissible. Protection and preservation is therefore very important. This is achieved through hashing the trace data. The evidence file should be copied; analysis should be done on the forensic copy and not the original file (Setup and Test Procedures - NIST, 2002).

Examination obtains traces from various security sensors that are integrated and fused to form one large dataset. Data is mapped, time lined and recovered from hidden files. Mapping and time lining of this data is also performed. Redundant information and unrelated data is removed, and minimum representative attributes are identified. The best

practice is to use the least information possible with the highest probable evidence (Franceschi & De, 2007).

In analysis, evidence collected is searched methodically to extract specific indicators of the crime. Indicators are classified and correlated to deduce important observations using the existing attack patterns. Important parameters include: network connection, establishment, DNS queries, packet fragmentation, protocol and operating system fingerprinting, running rogue processes and installed software or rootkits (Lyle, 2015).

Correlation and attribution are the hardest part of the investigation process. The information obtained from the evidence traces is used to identify who, what, where, when, how and why of the incident. It is meant for source traceback, reconstruction of the attack scenario and attribution to a source. Identifying the identity of an attacker is challenging due to IP spoofing or stepping stone attack (Adsit, 2016).

Presentation and review entail presenting observations in an understandable language to the organisation's management and legal personnel while providing an explanation of the various standard procedures used to arrive at a conclusion. Visual reports are easier to grasp and so they are the preferred method for documentation (Almulhem, 2016).

2.3. Web Crawlers

A web crawler is an application that visits web sites and reads their pages and other information to create entries for a search engine index (Monika, 2015). It can also be an automated script which browses the World Wide Web in a methodical, automated manner. A web crawler uses the concept of browsing the web page. Step one involves visiting the home page and then visit the links on the homepage.

Common use cases for crawlers are search engines. They use crawlers to update data from indexed websites. Web crawlers are mainly used to create a copy of all the visited pages for later processing by a search engine, that will index the downloaded pages to provide fast searches. Crawlers can also be used for automating maintenance tasks on a Web site, such as checking links or validating HTML code. Crawlers are often used to gather

specific types of information from Web pages, such as harvesting e-mail addresses (usually for spam) (Web crawler, 2017).

There are good and bad crawlers. Good crawlers include search engines bots such as Googlebot, BingBot and Yandex Bot. Googlebot is one of the most popular web crawlers on the Internet today as it is used to index content for Google's search engine. However, there are also issues sometimes when it comes to scheduling and load as a crawler might be constantly polling your site. Google's web crawler gives a lot of tools and control over the process. And this is where a *robots.txt* file comes into play. For example, a user can tell Googlebot not to index a certain page (Jackson, 2017).

Traditional, web crawlers retrieve all the pages that match the query, whether they are relevant for the user or not. In this era of big data, there is a need to prioritise our crawl and search. Due to limited time and resources available there is a need to develop efficient crawlers (Pal, Deepak, & Shrivastava, 2009). Focused crawlers narrow their scope to retrieve only relevant information for the user and discard the information which is irrelevant.

The process of extracting content from websites is known as web scraping. A powerful tool that can automatically extract information from web pages is grep command or regular expression-matching facilities of programming languages (Web scraping, 2017). A regular expression (regex or regexp for short) is a special text string for describing a search pattern. For example, text file wildcard notation is *.txt to find all text files in a file manager (Regular expressions.info, 2017). The regex equivalent for *.txt is ^. *\.txt$. The regular expression for Bitcoin address [13][a-km-zA-HJ-NP-Z0-9]{26,33}$.

2.4. Current Forensic Tools

Bitcoins have artefacts that can be used in a forensic investigation (Doran, 2015). Forensic tools help investigators to obtain evidence. There is minimal research done on cryptocurrency technology because it is relatively new. A few tools have been developed to solve the cryptocurrency forensics challenge. The Blockchain has a lot of information on Bitcoin transactions. There are blockchain analytic tools which have been developed

to decipher transactions. This section will review applications and services for forensic bitcoin transaction and blockchain analysis.

Chainalysis seeks to protect the integrity of digital assets (Chainalysis, 2017). It provides anti-money laundering software for Bitcoin. Hailed as an industry top player, Chainalysis accurately traced the history of bitcoin payments and wallets. It successfully mapped wallets into known clusters—that is, mapping addresses to known entities like Silk Road, Coinbase, and other large Bitcoin players (Farivar & Mullin, 2016).

Numisight provides a set of tools that are currently is public alpha (Numisight, 2017). The tools should verify all levels of every transaction on the Bitcoin blockchain. Though it is currently at the experimental stage it is designed to provide a visualization of bitcoin data and establish connections between transactions

Elliptic's technology prides itself in being the most timely and accurate in uncovering suspicious activities (Elliptic, 2017). Elliptic identifies illicit activity on the Bitcoin blockchain becoming a valuable tool to law enforcement agencies and banks (Buntinx, 2017). They have a proprietary database which provides auditable proof of identity of millions of Bitcoin addresses across real world entities.

Skry uses blockchain technology, big data and artificial intelligence to detect suspicious activities on the Bitcoin blockchain. Suspicious activities can be viewed publicly on the blockchain. The challenge comes in picking them real time. Skry uses artificial intelligence to allow for timely detection. The tools can help identify potential risks on the Bitcoin network. The intelligence blockchain platform helps identify non-compliance with AML, identify fraudulent patterns and predict unseen cyber-attacks (Skry Blockchain Intelligence Platform, 2017).

Another source of evidence is the host machine that has carried out a Bitcoin transaction. Devices used in Bitcoin transactions have a lot of evidentiary artefacts. By understanding the Bitcoin protocol, how it functions and the forensic artefacts on a network; forensic investigators can deal with the challenge of traceability and attribution. Network forensic tools such as NetworkMiner, and Wireshark can intercept transactions on a network.

NetworkMiner analyses layer 7 protocols and can identify Bitcoin on protocol 8333 (Pluskal, et al., 2015).

Forensic tools must attain a standard of accuracy for the evidence obtained to be considered admissible. It is imperative for a grading system of the forensic tools to ensure they are forensically fit for investigations. The goal of law enforcement communities such as the Computer Forensic Tool Testing (CFTT) project at the National Institute of Standards and Technology (NIST) is to establish a methodology for testing computer forensic software tools by development of general tool specifications, test procedures, test criteria, test sets, and test hardware (Computer Forensics Tools Testing Program, 2015). The results provide the information necessary for toolmakers to improve tools, for users to make informed choices about acquiring and using computer forensics tools, and for interested parties to understand the tools capabilities.

2.5. Conclusions

The Bitcoin network continues to grow. The average amount transferred per Bitcoin transaction is larger than in any other major payment network. The number of merchants accepting Bitcoin for trade is also on the rise. The identity of the user behind a Bitcoin address remains unknown until information is revealed during a purchase or in other circumstances.

Blockchain technology is very secure, it ensures the irreversibility of transactions. Secure, low cost of transactions and convenience are some of the reasons Bitcoin continues to grow. Despite the progress achieved in Bitcoin adoption, Bitcoin has not been regulated and is only accepted as a digital asset. This can be attributed to the low level of awareness and potential loss of control by governments and financial institutions. One of the major concerns for Bitcoin opponents is its preferential use by terrorists and its role in propagating cybercrimes. It is for these reasons that forensic investigations are crucial in tackling these challenges. Evidence presented in a court of law ought to be admissible,

complete, authentic and believable. The biggest challenge for Bitcoin Forensics is obtaining evidence in a timely manner.

Decentralised identity management makes it possible for a user to own an infinite number of identities and hence make it harder for investigators to nab criminals. Identities in Bitcoin are referred to as addresses. The Bitcoin address is hash of a user's public key has a unique format which 26-35 characters long. An address is constantly changed when one wants to reduce traceability. An address can reveal crucial information such as the number of outputs which can be used to infer real-life identities.

Gaps identified in literature review include lack of an address repository or database that can be queried by investigators when collecting data. The public ledger is one of the sources for evidence since it permanently records all transactions in the block. Building a Bitcoin case would entail obtaining possible identities used in crimes and observing their patterns in the blockchain. Bitcoin exchanges are obliged to cooperate with investigators in providing details about certain transactions.

Most of the existing Bitcoin forensic tools are still on the testing stage. The focus is on blockchain analysis processes. Most of the tools are not validated by CFTT. Chainalysis played a significant role in resolving the Silk Road crime ring. Bitcoin forensics brings about the importance of employing various leading technologies such as big data and artificial intelligence. This will be instrumental in resolving crimes around cryptocurrencies.

This research will focus on obtaining and archiving known addresses that can be used by investigators to infer and build a case in a court of law. Aim is to develop a cryptocurrency crawler that will regularly visit top cryptocurrency sites, parse, gather and store addresses with their website occurrences in the database using regular expressions. The stored data should provide as much information as possible to the investigators. Obtaining addresses and other Bitcoin artefacts is a challenge because identities can always be changed. It is crucial for constant extraction to be carried to ensure that the repository has as much information as possible. There is a skills gap in tacking Bitcoin forensics. More research

and training on the Bitcoin technology are long term solutions that law enforcement agencies should pursue aggressively.

Chapter 3: Methodology

3.1. Introduction

Rapid Application Development (RAD) will be used to achieve some of the outlined research objectives; it is a type of incremental model. Key objective is for fast development in the limited time frame and development of a high-quality system at a relatively investment cost (CMS, 2005). Project control involves prioritising development and defining delivery deadlines or "timeboxes". If the project starts to slip, the emphasis is on reducing requirements to fit timebox, not in increasing the deadline. It iteratively produces a production software. The RAD model has four phases: requirements planning, user design, development, and cutover. There is a cycle between the user design and construction phases (Introductions to RAD, 2017).

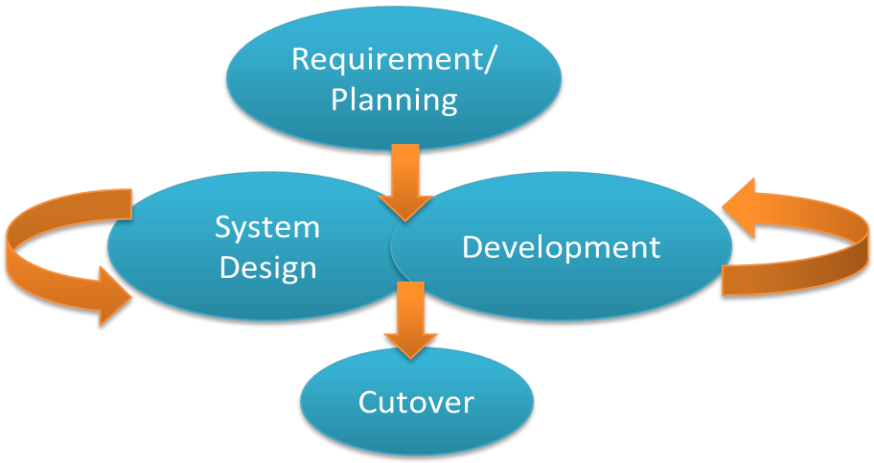


Figure 0-1: RAD Methodology

Advantages of RAD is that it concentrates on essential system elements from the user viewpoint. It also provides the ability to rapidly change system design as demanded by users. The application is to be developed in a short time frame; RAD is the right methodology for fast development and prototyping. Splitting the projects into modules will help in determining the pace of achievements (Introductions to RAD, 2017). Object oriented analysis and design will skillfully assign responsibilities to software components.

RAD is an iterative framework. Iterations for this project will be four weeks. The proposed timeline of the project is in the Appendix A. Each iteration goes through the process of requirement gathering to the end of the cycle.

3.2. Requirement Planning

The process will involve identifying business needs, project scope, constraints, and system requirements. Requirement gathering will be from secondary sources such as financial crimes journal and law enforcement agencies such as Federal Bureau of Investigation and case laws relating to Bitcoin crimes. Information on law enforcement guidelines will assist in meeting the compliance requirements.

A feasibility study will be used to check that the requirements are specific and attainable. After requirement gathering, requirements will be analysed for their validity, and the possibility of incorporating the requirements in the system to be development will be outweighed. Requirements specifications will be documented for the next phase.

3.3. System Design

On the design aspect, requirements are consolidated into modules. The system and software design will be prepared from the requirement specifications which were studied in the first phase. System Design helps in specifying hardware and system requirements and helps in defining overall system architecture. User requirements identified in requirement analysis will be used in system design specification.

Functional specifications will be documented graphically. Deliverables will be context diagrams which define the scope of the system under study. A logical data flow diagram of the proposed system should answer questions such as who will perform the task, how the tasks will be performed and the media type (Sangolly, 1997).

Computer-aided software engineering (CASE) tools will be used to develop software that is high-quality, defect-free, and maintainable (Computer Aided Software Engineering, 2016). Case tools to be used in system development Microsoft Visio, Microsoft Project and Microsoft Visual Studio Integrated Development Environment.

1. Microsoft Visio will be used to prepare UML diagrams (Microsoft Visio 2016, 2016).
2. Microsoft Project: is designed to assist in developing a plan, assigning resources to tasks, tracking progress, managing the budget, and analysing workloads (Microsoft Project, 2017).
3. Microsoft Visual Studio: is an integrated development environment (IDE) from Microsoft. It has a collection of CASE tools which include built-in tools include a forms designer for building GUI applications, web designer, class designer, and database schema designer (Visual Studio, 2017).

The deliverables are used in the development phase.

3.4. Development

System design documents are essential in this phase. Development focuses on program and application development tasks. Users continue suggests changes or improvements as actual screens or reports are developed. The system will be divided into modules/units and actual coding starts. The focus of this phase is the code. Most of the time will be allocated to the software development life cycle.

The creation of the software is well designed, robust, and maintainable software using object technologies and language such as C++ or C# (Visual Studio, 2017). Programming will be done in the following development in an integrated development environment (IDE) that is suitable for meeting the objectives.

1. Windows 10 Operating System
2. Microsoft Visual Studio IDE (C# programming language)
3. Xampp: Apache server and MySQL Database (XAMPP Apache + MariaDB + PHP + Perl, 2016).

Secure software development techniques will be used to develop the system. The objects defined under design will be mapped into code.

3.5. Cutover

Processes involved in this phase include testing and changing over to the new system.

3.5.1. Testing

Three types of testing will be done: functional, non-functional and structural testing. Testing is done iteratively on each prototype.

3.5.1.2. Functional Testing

Functionality testing is used to check that a software application performs and functions correctly per design specifications (What is Functionality testing? 2016). Core application functions, text input, menu functions and installation, and setup on localised machines, etc. are tested.

The following are the fundamental tests to be done in functionality testing:

1. Installation and setup on localised machines running localised operating systems and local code pages.
2. Check for core application functions: website crawling, address parsing, storage and archival of accurate.
3. String handling, text, and data, especially when interfacing with non-Unicode applications or modules.
4. Text handling (such as copying, pasting, and editing) of extended characters, special fonts, and non-Latin scripts.

3.5.1.3. Non-functional requirements tests

1. Non-functional tests such as stress and performance testing will be carried out by increasing the maximum number of sites that can be crawled.
2. Security testing with respect to authentication and hacking.
3. Quality assurance: the source code and data flow diagrams will be evaluated to see if they satisfy the logical requirements.

3.5.1.4. Structural Testing

Structural testing involves testing of the structure of the system or component. During structural testing, the tester is concentrating on how the software does it. For example, a structural technique wants to know how loops in the software are working. Various test

cases may be derived to exercise the loop once, twice, and many times. This may be done regardless of the functionality of the software.

It is critical for a forensic tool to be subjected to an expert's review. The logic of the code will affect the output. Forensic processes have no room error. The source code will be reviewed by an expert.

3.5.1.5. Validation

Application systems must be validated to ensure accuracy, reliability, consistent intended performance, and the ability to discern invalid or altered records. Digital Forensics is founded on the principles of repeatable processes and quality evidence (Brunty, Validation of Forensic Tools and Software: A Quick Guide for the Digital Forensic Examiner, 2011). Forensics requires that the validation process to be handled scientifically. The developed application will therefore be internally validated using four basic steps.

Step one involves developing a plan which will include doing a background check on the basic functionalities of forensic extraction tools using NIST guidelines. The basic functionalities include creating, hashing, and write blockers. The CFTT also publishes detailed validation reports on various types of forensic hardware and software on disk imaging tools. All the requirements outlined in requirement planning will be counter checked to validate the tool.

The next step will involve preparing a controlled data set which will be prepared test cases. The extraction tool will be used to extract the test data from a webpage (Lyle, 2015). This will not only assist in future validations, but may also be handy for internal competency and proficiency examinations for digital examiners. Step three will involve conducting a test in a controlled environment. It will involve using a repeatable, consistent, scientific method in drafting these validations

The last step is result validation. Results generated through the experimentation and validation stage must be repeatable. It is recommended that each requirement be tested at least three times. The output is compared with the expected results. If the results match, it means that the tool is valid and can be used for future investigations. Variables that may affect the outcome of the validation will be identified and listed. The question that will be

answered in this step is if the Bitcoin address database will assist investigators in gathering evidence.

3.5.2. Change Over to the New System

The system is not going to replace an existing system hence changing to the new system will be straightforward. After successful testing the product is deployed to the customer for their use. Beta testing involves sending the system to users who install it and use it under real-world working conditions. The testers of the system will be my technology students. If any changes are required or if any bugs are caught, then they will report it. Once those changes are made, or the bugs are fixed then the final deployment will happen.

The users to test this application will be a group conversant with the forensic process.

Chapter 4: Requirement Planning

4.1. Introduction

This phase is important in the software development lifecycle. All the requirements are described in this phase. Any changes made into the future must be incorporated in this document. Requirement specification will be used in the testing phase to confirm if the objectives are met. After using case studies and other secondary materials; the following requirements were identified.

This dissertation project aims at extracting Bitcoin addresses through focused crawling of webpages. The addresses are stored in a repository with their URL instances. Investigators can search the repository whenever they are building a case. All transactions related to the address can be checked on the blockchain and possible real-world identities inferred. The developed application should be robust and process (i.e., crawl) as many pages as possible.

The requirements are both functional and non-functional.

4.2. Functional Requirements

The major requirement is developing a focused crawler that visits cryptocurrency sites and extracts Bitcoin addresses. Modules consists of crawler, address extractor and a search feature.

1. Crawl multiple sites concurrently
2. Pause/resume live crawls
3. Simplified pluggability/extensibility
4. Avoid getting blocked by sites
5. Automatically tune speed/concurrency
6. Index pages
7. Parse web pages to extract Bitcoin addresses
8. Store Bitcoin addresses

4.3. Non-Functional Requirements

4.3.1. System Performance Requirements

1. Speed of crawling
2. Efficiency in terms of the accuracy of scraping data.
3. Reliability

4.3.2. Hardware Requirements

Machine:	Intel Core i5 or higher
Clock Speed:	2.5 GHz or higher
System Memory:	4 GB

4.3.3. Software Requirements

Operating Systems:	Windows 10
Web Server:	Apache Server 2.4.23
Front – end:	C#
Development Platform:	.Net Framework 4.5 or later

Chapter 5: System Design

5.1. Proposed System Modules

Design is the first step in the development process. The goal is to design a model of an entity that will be built. Quality of the system is fostered in this phase. A strong design translates into a stable system that can be tested.

The proposed system is a cryptocurrency focused crawler. The goal of a focused crawler is to selectively seek out webpages that are relevant for the crawler. The proposed system architecture has four main modules: *Crawler, Html-Converter, Address Scraper* and *Searcher*.

Crawler: This module will visit webpages from a list of URLs in a plain-text document. It also visits the hyperlinks in the webpages and linked documents (to depth X). The HTML pages are downloaded for the next process.

Html-Converter: The HTML pages are converted into text documents that can be easily parsed to extract the target data.

Address Parser: This module involves using regular expressions to extract Bitcoin addresses on the resulting text document from *Html-Converter*.

Searcher: This module is used by investigators looking for evidence of a certain address. The user can give his/her query under any specified category and best matched results of the query from that category will be retrieved. It could be 1:1 match.

5.2. Logical System Design

Logical design describes the system and the process by which is developed. It refers to the technical specifications that will be applied in implementing the proposed system. It also includes construction of program and designing of output, input, code, database and process of the system.

5.2.1. Use-case Diagram

A use-case diagram models the behaviour of the system, a subsystem or class. It is more important for visualising, specifying and making systems and subsystems approachable.

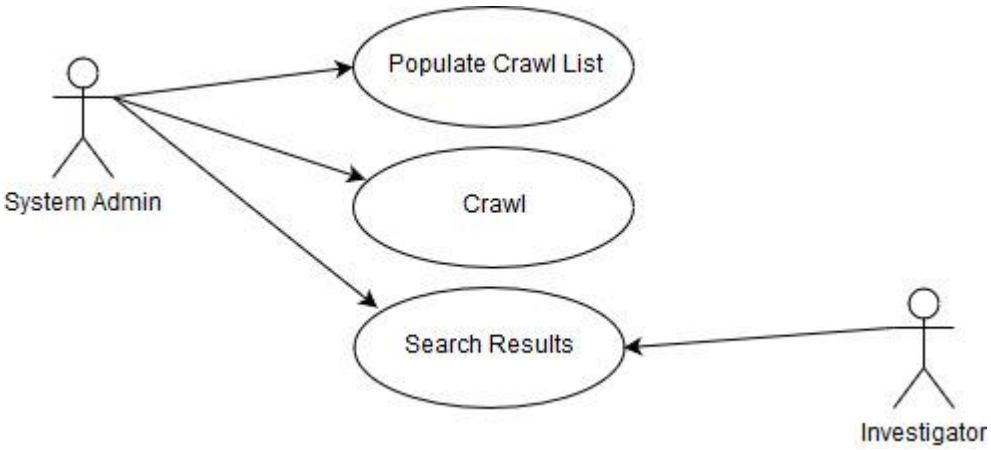


Figure 0-1: Use-case Diagram

The Use-case diagram in Figure 5-2, has two actors the Investigator and System admin. The use cases System admin is to populate the Crawl list., *Crawler* involves scheduling and initiating the crawl. In the *Search Results* use case, both the investigator and system admin can query the results of the crawl.

5.2.2. Flowchart

A flowchart is a diagram that represents an algorithm, workflow or process, showing the steps as boxes of various kinds, and their order by connecting them with arrows.

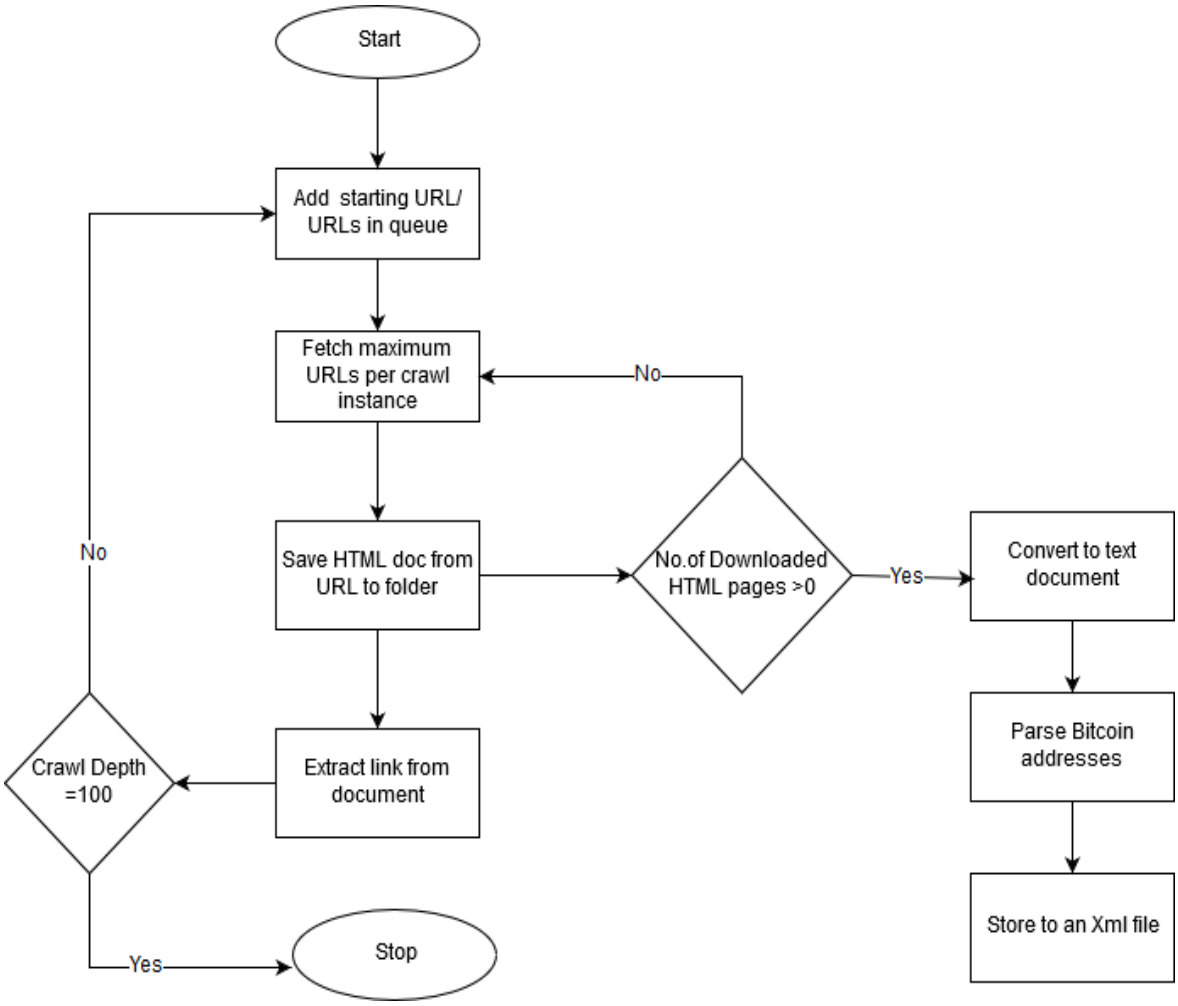


Figure 0-2: Flowchart

The Flowchart in Figure 5-3 explains the basic flow of Cryptocurrency Crawler. The crawler is initialised with some URLs depending on the (maxConcurrentSiteCrawls) specified in the configuration file. The maximum crawl depth is 100 pages; this determines the internal links to be added in the queue. This process continues until a stopping condition is reached or there is no more webpage's in the URL Queue to crawl. Maximum concurrent site crawls and the crawl depth can be adjusted by the system administrator.

5.3. Physical System Design

The physical system specifies the design specifications that defines what is to be programmed. Specifications are supposed to meet input, output, storage, processing requirements. In this stage, physical design will consider user interface design and data design (System Design, 2017).

5.4. User Interface Design

User Interface Design deals with how users add information to the system and with how the system presents information back to them. The Crawler is populated with URLs from a text document maintained by the system administrator. The developed system is a console application that uses log4net to display the progress of the crawl. Crawled pages can either be online or on the local server. A proper connection to the apache server is needed to access webpages on the localhost. Log4net framework is based on Apache log4j (Logging Services, 2017). The application saves parsed results in an XML file.

5.5. Data Design

Data Design deals with data representation and storage within the system.

5.5.1. Entity Relationship Diagram

An entity-relationship diagram (ERD) depicts an information system and the relationship between people, objects, places, concepts or events within that system. It is an abstract way of describing the database.

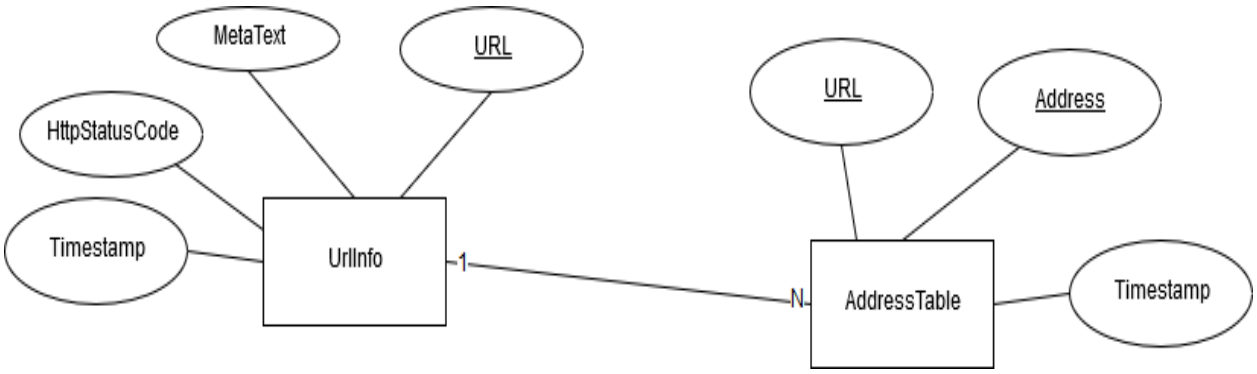


Figure 0-3: ERD Diagram

ER Diagram in Figure 5-1, in this diagram there are two entities:

Entity 1: AddressTable attributes: the AddressTable stores the scraped addresses and its metadata. The time it was retrieved and on which site.

URL: the URL which address was parsed.

Address: It is a 26-35 alphanumeric characters' identifier, beginning with the number 1 or 3, that represents a possible destination for a Bitcoin payment. It is the keyword that will be parsed from webpages.

Timestamp: used to define at which moment in time an address was added or updated.

Entity 2: UrlInfo attributes: stores Urls that have been crawled by the developed application. Records the time the URL was crawled and the status code.

URL: Complete URL of the webpage.

MetaText: gives a description of the webpage.

Http Status Code: Indicate whether a specific HTTP request has been successfully completed.

Timestamp: used to define at which moment in time a URL was added or updated.

The relationship of *UrlInfo* to *AddressTable* is one to many. A page can have many addresses.

Schema

Figure 0-4: Database Schema

Output is serialised to an XML document.

DTD Schema:

AddressTable:

```
<!DOCTYPE AddressTable
[
<!ELEMENT AddressTable (Address,URL,Timestamp)>
<!ELEMENT Address (#PCDATA)>
<!ELEMENT URL (#PCDATA)>
<!ELEMENT Timestamp (#PCDATA)>
]>
```

UrlInfo :

```
[
<!DOCTYPE UrlInfo
[
<!ELEMENT UrlInfo (URL,HttpStatusCode,MetaText,Timestamp)>
<!ELEMENT URL (#PCDATA)>
<!ELEMENT HttpStatusCode (#PCDATA)>
<!ELEMENT MetaText (#PCDATA)><!ELEMENT Timestamp (#PCDATA)>
]>

]>
```

Chapter 6: Development

Development involves using the architectural design and the results of system analysis to construct system elements that meet the outlined system requirements developed in the design phase.

The application is built on AbotX (AbotX, 2017), an open source web crawler that provides set of wrappers and extensions. It allows for concurrent crawling of multiple sites, auto tuning and auto throttling (Mourier, 2015).

Auto tuning automatically monitors the host machine's resource usage and adjusts the crawl speed and concurrency to maximise throughput without overrunning it. Auto throttling automatically slows down the crawl speed if the website being crawled is showing signs of stress or unwillingness to respond to the frequency of http requests. Parameters on the crawl include maximum sites, crawl depth and crawl timeout seconds.

Input data: The crawler crawls and scrapes URLs from an external source Sites.txt. This list is updated by the system administrator. Sites must be in http format, see Appendix C.

Crawler Configuration:

Configurations values can be tweaked depending on the system performance of the machine.

```
</abotX>
<abotX maxConcurrentSiteCrawls="3"
sitesToCrawlBatchSizePerRequest="25"
minSiteToCrawlRequestDelayInSecs="15"
isJavascriptRenderingEnabled="false"
javascriptRenderingWaitTimeInMilliseconds="3500"> //configs
for crawler
    <autoThrottling isEnabled="false" thresholdMed="5"
thresholdHigh="10" thresholdTimeInMilliseconds="5000"
minAdjustmentWaitTimeInSecs="30" /> // slows down the crawl
speed if the website
    <autoTuning isEnabled="false" cpuThresholdMed="65"
cpuThresholdHigh="85" minAdjustmentWaitTimeInSecs="30" />
//automatically monitors the host machine's resource usage
and adjusts the crawl speed and concurrency to maximize
throughput without overrunning it
    <accelerator concurrentSiteCrawlsIncrement="2"
concurrentRequestIncrement="2"
delayDecrementInMilliseconds="2000"
minDelayInMilliseconds="0" concurrentRequestMax="10"
concurrentSiteCrawlsMax="3" /> // increment the number of
sites per crawl
    <decelerator concurrentSiteCrawlsDecrement="2"
concurrentRequestDecrement="2"
delayIncrementInMilliseconds="2000"
maxDelayInMilliseconds="15000" concurrentRequestMin="1"
concurrentSiteCrawlsMin="1" /> // decrement the number of
sites per crawl
  </abotX>
```

Populate Crawler:

```
var fileStream = new FileStream("Sites.txt", FileMode.Open,
FileAccess.Read);//open Sites.txt to read websites
            using (var streamReader = new
StreamReader(fileStream, Encoding.UTF8))
            {
                string line;
                while ((line = streamReader.ReadLine()) !=
null)
                {
```

```
                    siteToCrawlProvider.AddSitesToCrawl(new
List<SiteToCrawl>
                {
                    new SiteToCrawl {Uri = new Uri(line)
                }
                });// add sites to a list

                    Console.WriteLine(line);//display the
read sites
```

Sample Crawl:

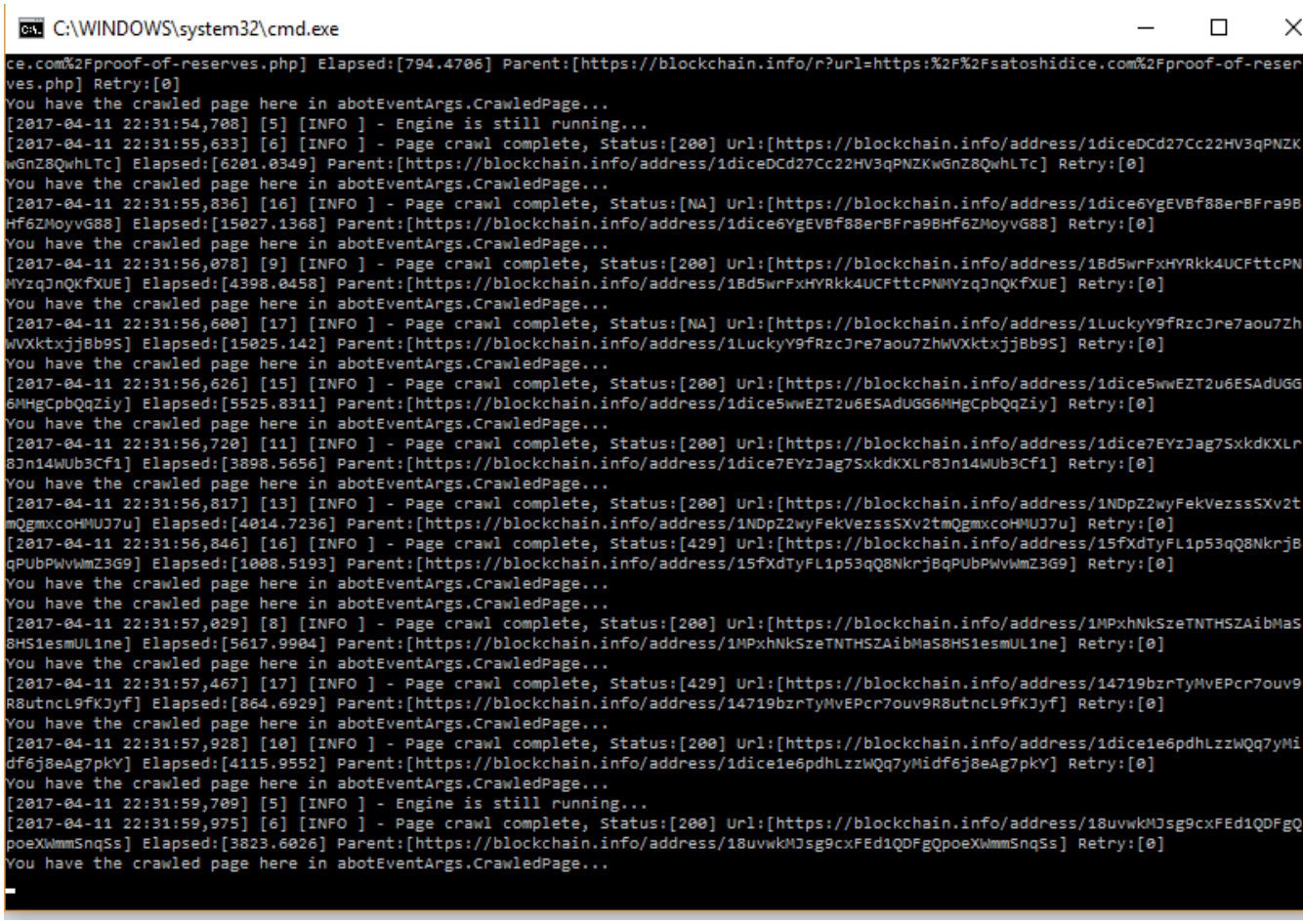


Figure 0-1: Sample Crawl

Expected output from webpages with Bitcoin addresses is Bitcoin addresses. Sample output from this crawl is displayed in Appendix B:

Chapter 7: Cutover

7.1. Testing

This step involves testing and validation. The program is run to test for errors

Testing Objectives:

1. To find errors
2. To enhance integrity of a system by detecting deviations in the design
3. To detect error prone areas
4. To ensure customer requirements are met

Test case description:

Test cases were used to validate the developed tool. They consist of webpages with Bitcoin addresses which were downloaded and stored in a local webserver. Validation involved confirming the addresses were identified and stored in the repository. The expected results are compared to the actual results.

7.1.2. Functional Testing

Functional testing checks if core requirements and functionalities are met.

1. *Crawl multiple sites concurrently*: the crawler crawls multiple sites concurrently. An administrator can increase or reduce number of sites per crawl depending on the performance of the site.
2. *Create logs*: logs are stored in a text file
3. *Input*: the system administrator populates a text document that serves as the input for the crawler.
4. *Pause/resume live crawls*: this is achieved through the following code.

```
System.Threading.Thread.Sleep(3000);
            crawlEngine.Pause();
            System.Threading.Thread.Sleep(10000);
            crawlEngine.Resume();
```

5. Index pages: pages are assigned an index for easy retrieval

```
config.ConfigurationExtensions.Add("Somekey1",
"SomeValue1");
```

6. *Parse web pages to extract Bitcoin addresses*: the application uses Bitcoin's regular expression to match addresses on webpages.

```
Regex g = new Regex("[13][a-km-zA-HJ-NP-Z0-
9]{26,33}");// regular expression for a Bitcoin
address
```

7. *Store Bitcoin addresses*: extracted addresses and its metadata are stored in an xml file.

```
XmlSerializer serial = new XmlSerializer(typeof(List<Addresses>));//
serialize addresses to xml file
                                System.IO.StreamWriter writer = new
System.IO.StreamWriter(@"C:\files\address.xml");
                                serial.Serialize(writer, eCrawlResult);
                                writer.Close();
```

7.1.3. Non-functional Testing

Parameters tested included the performance of the application.

Crawlers can retrieve data much quicker and in greater depth than human searchers, so they can have a crippling impact on the performance of a site. If a single crawler is performing multiple requests per second and/or downloading large files, a server would have a hard time keeping up with requests from multiple crawlers. The *app.config* file defines parameters to be used in the crawl. Parameters such as maximum concurrent crawls, crawl depth and crawl timeout allow for application optimisation.

By converting HTML to a text document, the Bitcoin address regular expressions successfully matches addresses in the document. Figure 7-1 details the progress of a crawl for blockchain.info

7.1.4. Structural Testing

In structural testing, tests are derived from the knowledge of the software's structure or internal implementation. Structural testing is critical because the output of the tool is a

meant for the forensic process. The use of an existing framework helped in reducing the number of errors because the bugs had been identified and improved on. The development process involved constant peer review to counter check the logic of the code.

7.2. Validation

Key to the forensic process is results that are repeatable and quality evidence. NIST's guidelines state that the test results of forensic software or tools should be repeatable and reproducible (Brunty, Validation of Forensic Tools and Software: A Quick Guide for the Digital Forensic Examiner, 2017).

Repeatability refers to obtaining the same results when using the same method on identical test items in the same laboratory by the same operator using the same equipment within short intervals of time. The crawl was performed four times on the test cases and produced the same results.

Reproducibility refers to obtaining the same results being obtained when using the same method on identical test items in different laboratories with different operators utilising different equipment. The application was installed on another machine and peer reviewed using the same test case and it resulted in the same results.

Test Case: 100 Most Popular Addresses as per Blockchain.info

The crawler should extract 100 Bitcoin addresses listed on the blockchain.info. Expected output is 100 Bitcoin addresses

URL: https://blockchain.info/popular-addresses

Blockchain.info popular addresses website screenshot:

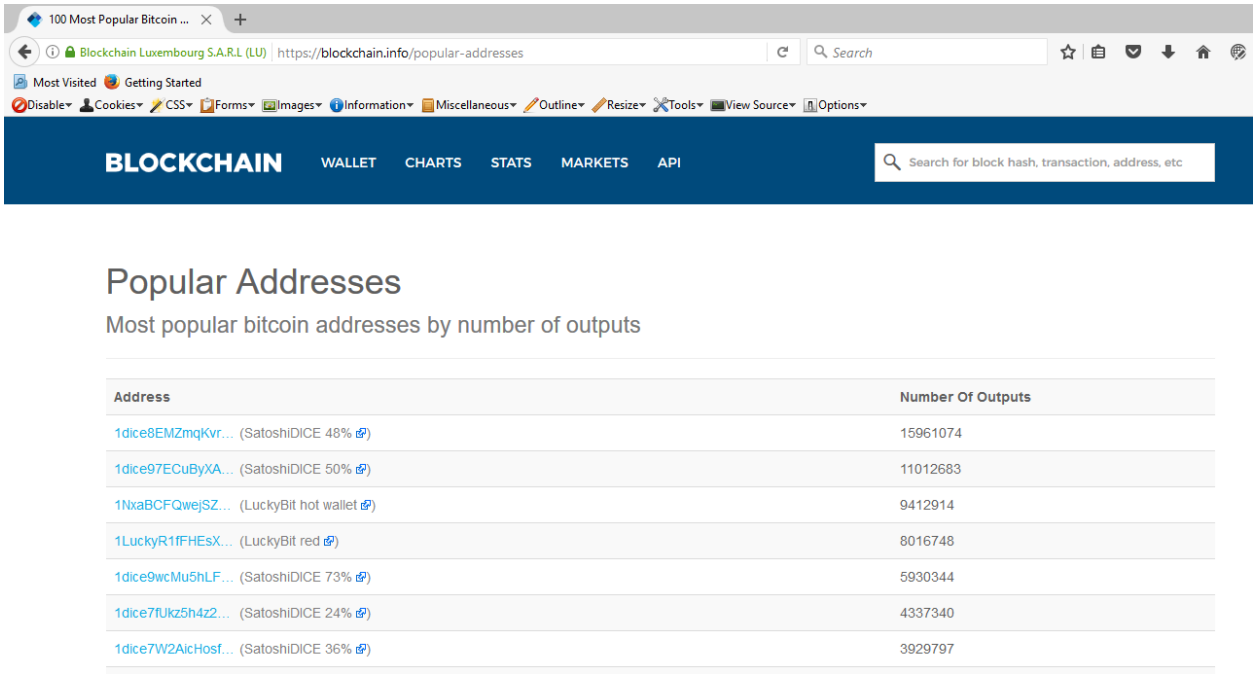


Figure 0-1: Test Case Blockchain.info/Popular-Addresses

Crawl Progress:

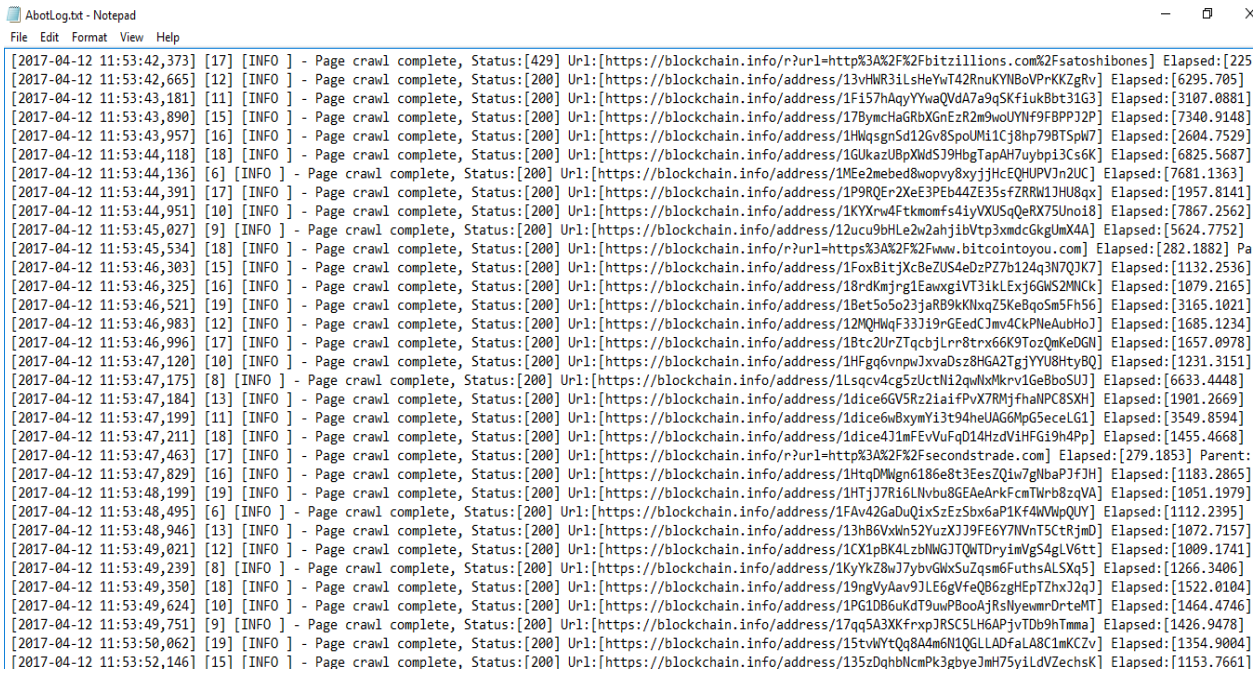


Figure 0-2: Crawl Progress

Regular expression source code to parse Bitcoin Addresses:

```
MatchCollection addresses = Regex.Matches(line, search);//
uses matchcollection to store Bitcoin addresses
                    List<Addresses> eCrawlResult = new
List<Addresses>();

                        if (addresses.Count > 0) //Count
No.of addresses
                        {
                            for (int a = 0; a <
addresses.Count; a++)
                            {


Console.WriteLine(addresses[a] + " " + "is Address No." + "
" + a);// Display parsed addresses
                                    // Y.
                                    // Write original line and
the value.
                                    string v =
m.Groups[1].Value;
```

100 popular addresses parsed from https://blockchain.info/popular-addresses

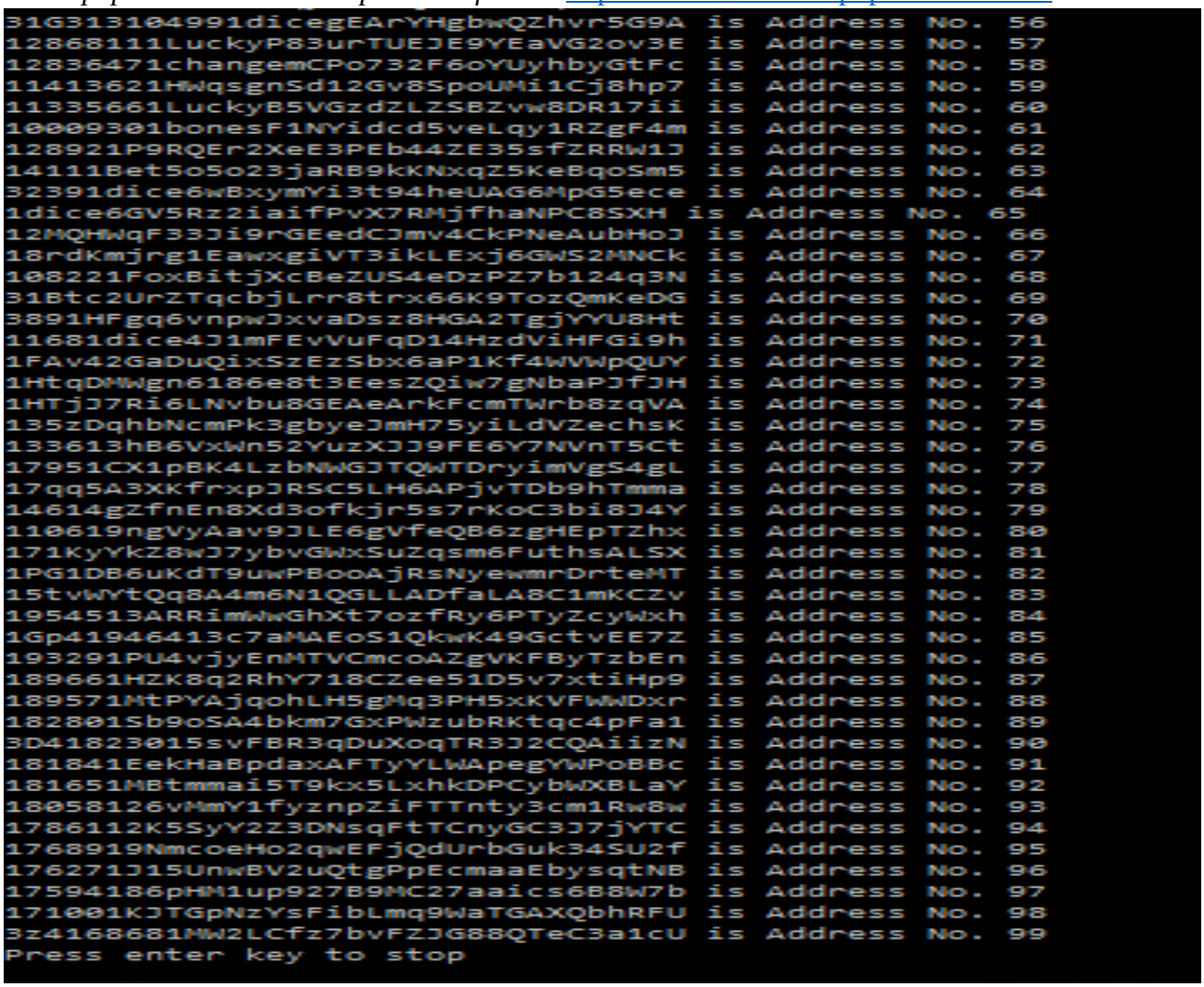


Figure 0-3: Parsed Addresses from Blockchain.info

The application successfully parses all addresses from https://blockchain.info/popular-addresses. Figure 7-3 shows all the addresses parsed from the website.

7.3. Change Over to the New System

The framework (AbotX) used in the development of this application is highly extensible and very well written. It is also well managed, and there are regular updates to the code base. This makes it easy for it to be transitioned into a production environment. There is

an option to tweak how the crawler appears as a client, to respect robots, and inject handlers with the ability to extend built in classes.

Chapter 8: Discussion of Results

This research identified that Bitcoin payments are highly preferred by criminals such as terrorists and cyber criminals. This is because they offer anonymity when conducting transactions online. Personally, identifiable information is not needed to make a transaction. It is for this reason that forensics is important to prosecute criminals. A public ledger known as blockchain records all Bitcoin transactions.

Bitcoin forensics is inhibited by lack of digital evidence needed in a court of law. A user can use multiple identities that cannot be linked to his real-world identity. This research proposed a Bitcoin address repository that would store addresses from cryptocurrency sites. The developed application uses regular expressions to match and parse Bitcoin addresses from webpages listed by the system administrator.

Chapter 9: Conclusions, Recommendations and Future Work

9.1. Conclusions

The use of cryptocurrencies is on the rise; Bitcoin is one of the most used cryptocurrencies. Bitcoin transactions allow for infinite identities when making payment. This makes it difficult for investigators to find culprits when Bitcoins are used in crime funding. To solve the challenge of collecting Bitcoin transactions related evidence, an address repository was designed. This is achieved through crawling targeted cryptocurrency sites. The developed application extracts Bitcoin addresses from the crawled webpages. An investigator can search an address on the repository and check for its transaction data on the blockchain. The information in the repository will assist an investigator to build a case. Metadata of the addresses such as the URL which the address was extracted helps an investigator identify services that used the Bitcoin address. The blockchain shows all the amount send to that the address, sender's address and time.

Decentralised technologies are relatively new, there is a skills gap in investigating cryptocurrency related crimes. Most jurisdictions have not formulated policies to be implemented when dealing with cryptocurrencies. This is a challenge for the legal systems working on virtual currencies related cases.

9.2. Recommendations

Recommendations for this project include the use of the developed tool for large scale parsing of cryptocurrency sites. In a server environment, the crawl can be adjusted to match the resources available. The larger the crawl depth the richer the Bitcoin address repository. Crawling should be scheduled regularly to ensure timely capture of addresses from websites.

Education around cryptocurrencies would help law enforcement agencies conduct successful investigations. Besides the law enforcement agencies, the magistrates are also not well versed with the technology which makes it hard to determine a case. It is important for school curriculums to include courses on cryptocurrencies. A legal framework governing the use of cryptocurrencies can be worked on using joint efforts from law enforcement agencies, governments and financial institutions.

Another critical aspect is the validation of forensic tools developed. Bitcoin forensic tools are relatively new and most are still on the testing stage. Speedy validation by the various bodies would help in the investigation process. Evidence from the validated tools would be considered admissible in a court of law.

9.3. Future Work

Parsing content from webpages is a resource intensive process, more research can be done on less intensive methods of parsing webpages. Some forms of crawling are intrusive and end up changing the site content. Focused crawling would also be beneficial in optimising resource use because only relevant pages would be downloaded. The blockchain supplements the address repository because transaction data associated with the addresses can be found on the blockchain.

The dark and deep web is where most fraudulent activities such as the Silk Road market place. Crawling sites on the deep web would result in an information rich repository. The application can be designed to visit pages on the dark web. Indexing of the web pages can assist in determining will assist in informing the crawling decisions by the administrators.

More research should be done around forensics for alternative coins such as Ethereum, Monero and Dash. The application can also be linked to the blockchain to allow for quick search of transaction data of collected addresses. The Bitcoin blockchain has a vast amount of information that can be utilised in forensic investigations. Tools that can perform blockchain transactions analysis real time are on high demand. Timely capture of suspicious activities can be achieved through artificial intelligence and big data.

References

AbotX. (2017, 4 12). Retrieved from abotx.org: https://abotx.org/

Adsit, B. (2016, April 5). *Forensics (Xplico)*. Retrieved from Prezi: https://prezi.com/5xsvtkisiyav/forensics-xplico/

Almulhem, A. (2016, September 20). *Network Forensics : Notions and Challenges.* Retrieved from IEEEXplore: http://ieeexplore.ieee.org/document/5407485/?arnumber=5407485

Arvind, N. J., Edward, F., Andrew, M., & Steven, G. (2015). *Bitcoin and Cryptocurrency Technologies.* New Jersey: Princeton.

Bitcoin. (2016, December 3). Retrieved from Coinbase: https://www.coinbase.com/charts?locale=en

Bitcoin Developer Guide. (2017, March 29). Retrieved from Bitcoin: https://bitcoin.org/en/developer-guide#p2p-network

Bitcoin Developer Reference. (2016, December 7). Retrieved from Bitcoin Developer Reference: https://bitcoin.org/en/developer-reference#rpc-quick-reference

Bitcoin Miner. (2015, December 9). Retrieved from Bitcoin Miner: http://bitcoinminer.com/

Bitcoinwiki : Address. (2017, March 20). Retrieved from Bitcoin Wiki: https://en.bitcoin.it/wiki/Address

Bitcoinwiki. (2016, Septemeber 18). Retrieved from Wiki: https://en.bitcoin.it/wiki/Protocol_documentation

Bitcoinwiki: Protocol documentation. (2016, September 18). Retrieved from Wiki: https://en.bitcoin.it/wiki/Protocol_documentation

Block Explorer. (2016, December 7). Retrieved from Block Explorer: https://blockexplorer.com/

Blockchain. (2016, June 6). Retrieved from Blockchain: https://www.blockchain.com/

Blockchain.Info - Popular Addresses. (2017, July 2). Retrieved from Blockchain.Info: https://blockchain.info/popular-addresses

Brill, A., & Keene, L. (2014). Cryptocurrencies: The Next Generation of Terrorist Financing? *Defence Against Terrorism Review*, Vol. 6, No. 1 pp. 7- 30.

Brunty, J. (2011). Validation of Forensic Tools and Software: A Quick Guide for the Digital Forensic Examiner. *Forensic Mag*.

Brunty, J. (2017, April 8). *Validation of Forensic Tools and Software: A Quick Guide for the Digital Forensic Examiner*. Retrieved from Forensic Mag: http://www.forensicmag.com/article/2011/03/validation-forensic-tools-and-software-quick-guide-digital-forensic-examiner

Buntinx, J. (2017, February 12). *Top 4 Companies Providing Bitcoin Blockchain Analysis Services.* Retrieved from The Merkle: https://themerkle.com/top-4-companies-providing-bitcoin-blockchain-analysis-services/

Chainalysis. (2017, June 3). Retrieved from Chainalysis: https://www.chainalysis.com/

CoinDesk. (2015, March 20). Retrieved from Coin Desk.com: http://www.coindesk.com/information/how-do-bitcoin-transactions-work/

Computer Aided Software Engineering. (2016, November 30). Retrieved from Wikipedia: https://en.wikipedia.org/wiki/Computer-aided_software_engineering

Computer Forenscis Tools Testing Program. (2015, August 15). Retrieved from NIST: http://www.cftt.nist.gov/

Cottle, M. (2016). The Government's Perilous Bitcoin Chase. *The Daily Beast*, 25 June.

Crypto-Currency Market Capitalizations. (2016, December 1). Retrieved from Coin Market: https://coinmarketcap.com/

Doran, M. (2015). Forensic Bitcoin. *SANS Institute*, 14-36.

Elliptic. (2017, June 3). Retrieved from https://www.elliptic.co/

Farivar, C., & Mullin, J. (2016, August 17). *Lost on the Silk Road*. Retrieved from ars Technica: https://arstechnica.com/tech-policy/2016/08/stealing-bitcoins-with-badges-how-silk-roads-dirty-cops-got-caught/

Franceschi, G. C., & De, A. (2007). *Open Source Network Forensic Analysis Tools*. Retrieved from Xplico: http://www.xplico.org/

Hashcash. (2016, December 14). Retrieved from Hashcash: http://www.hashcash.org/

Hunt, R. (2012, April 23). *New Developments In Network Forensics.* Retrieved from IEEE Xplore: http://doi.org/10.1109/ICON.2012.6506587

Introductions to RAD. (2017, February 25). Retrieved from http://www.ftms.edu.my/images/Document/IMM006%20-%20RAPID%20APPLICATION%20DEVELOPMENT/IMM006%20RAPID%20APPLICATION%20DEVELOPMENT%20-%20note%20chapter%201.pdf

Jackson, B. (2017, April 8). *Web Crawlers and User-Agents – Top 10 Most Popular*. Retrieved from keycdn.com: https://www.keycdn.com/blog/web-crawlers/

Logging Services. (2017, April 05). Retrieved from Apache.org: https://logging.apache.org/log4net/

Lyle, J. (2015). *Computer Forensics Tools Testing Program.* National Institute of Standards and Technology.

Microsoft Project. (2017). Retrieved from Microsoft: https://products.office.com/en-us/project/project-and-portfolio-management-software

Microsoft Visio 2016. (2016). Retrieved from Microsoft: https://www.microsoftstore.com/store/msusa/en_US/cat/Visio/categoryID.69407600

Milad, A., & Irwin, G. (2016). The use of crypto-currencies in funding violent jihad. *Journal of Money Laundering Control*, Vol. 19 ISS 4 pp - 1-22.

Monika, D. J. (2015). DESIGN OF CATEGORY-WISE FOCUSED WEB CRAWLER. *International Journal of Advance Research In Science And Engineering* , 1785-1787.

Mourier, S. (2015, April 7). *AbotX*. Retrieved from AbotX.org: maxConcurrentSiteCrawls

Nakamoto, S. (2016, August 31). *Bitcoin: A Peer-to-Peer Electronic Cash System.* Retrieved from Bitcoin: www.bitcoin.org

Nelson, B., Phillips, A., & Steuart, C. (2014, November 11). *Guide to Computer Forensics and Investigations*. Retrieved from Cengage Learning. isbn: 9781305176089.: https://books.google.cz/books?id=PUh9AwAAQBAJ

Numisight. (2017, June 3). Retrieved from Numisight: http://numisight.com/

Oxford Dictionary. (2017, February 7). Retrieved from Oxford Dictionary: https://en.oxforddictionaries.com/definition/artefact

Pal, A., Deepak, S., & Shrivastava, S. (2009). Efficient Focused Crawling Based on Content and Link Structure analysis. *International Journal of computer science and information security, vol. 2*.

Pluskal, J., Matoušek, P., Ryšavý, O., Kmeť, M., Veselý, V., Karpíšek, F., & Vymlátil, M. (2015). *Netfox Detective: A tool for advanced network forensic analysis.* Brno. Retrieved from http://www.fit.vutbr.cz/~matousp/pubs.php?file=%2Fpub%2F10863%2Fspi15v4.pdf&id=10863

Protect your privacy. (2016, December 2). Retrieved from Bitcoin: https://bitcoin.org/en/protect-your-privacy

Regular expressions.info. (2017, August 8). Retrieved from regular-expressions.info: http://www.regular-expressions.info/

Sangolly, J. (1997, October 2). *Systems Analysis & Design*. Retrieved from Albany.edu: http://www.albany.edu/acc/courses/fall97/acc681/ch7.html

Serialization (C#). (2017, April 7). Retrieved from Microsoft: https://msdn.microsoft.com/en-us/library/mt656716.aspx

Setup and Test Procedures - NIST. (2002, August 1). Retrieved from http://www.cftt.nist.gov/setup_for_dd_tests.pdf

Shusht, P., & Dostov, V. (2014). Cryptocurrencies: an unconvetional challenge to AML/CFT. *Journal of Financial Crime*, 239-243.

Silk Road (marketplace). (2017, February 4). Retrieved from Wiki: https://en.wikipedia.org/wiki/Silk_Road_(marketplace)

Simser, J. (2015). Bitcoin and the modern alchemy. *Journal of Financial Crime Vol.22 ISS 2*, 156-169.

Sira, R. (2003). *Network Forensic Analysis Tools.* SANS INstitute.

Smith, D. (2014, March 24). *3 Reasons The IRS Bitcoin Ruling Is Good For Bitcoin*. Retrieved from Nasdaq: http://www.nasdaq.com/article/3-reasons-the-irs-bitcoin-ruling-is-good-for-bitcoin-cm339333

System Design. (2017, April 6). Retrieved from Wikipedia.org: https://en.wikipedia.org/wiki/Systems_design

Tasca, P. (2015). *Digital Currencies: Principles, Tasks, Trends, Opportunities and Risks.* Zurich: Ecurex Research .

Technical background of version 1 Bitcoin addresses. (2016, January 29). Retrieved from bitcoinwiki: https://en.bitcoin.it/wiki/Technical_background_of_version_1_Bitcoin_addresses

Vesely, V. (2016, September 21). *Strathmore - Bitcoin Forensics*. Retrieved from Prezi: https://prezi.com/spci-vmya2lz/strathmore-bitcoin-forensics/

Visual Studio. (2017). Retrieved from Microsoft: https://www.visualstudio.com/

Web crawler. (2017, April 8). Retrieved from ScienceDaily: https://www.sciencedaily.com/terms/web_crawler.htm

Web scraping. (2017, April 8). Retrieved from wikipedia.ord: https://en.wikipedia.org/wiki/Web_scraping

What is Bitcoin Mining? (2017, January 31). Retrieved from bitcoinmining.com: https://www.bitcoinmining.com/

What is Functionality testing? (2016, December 30). Retrieved from ISTQB Exam Certification: http://istqbexamcertification.com/what-is-functionality-testing-in-software/

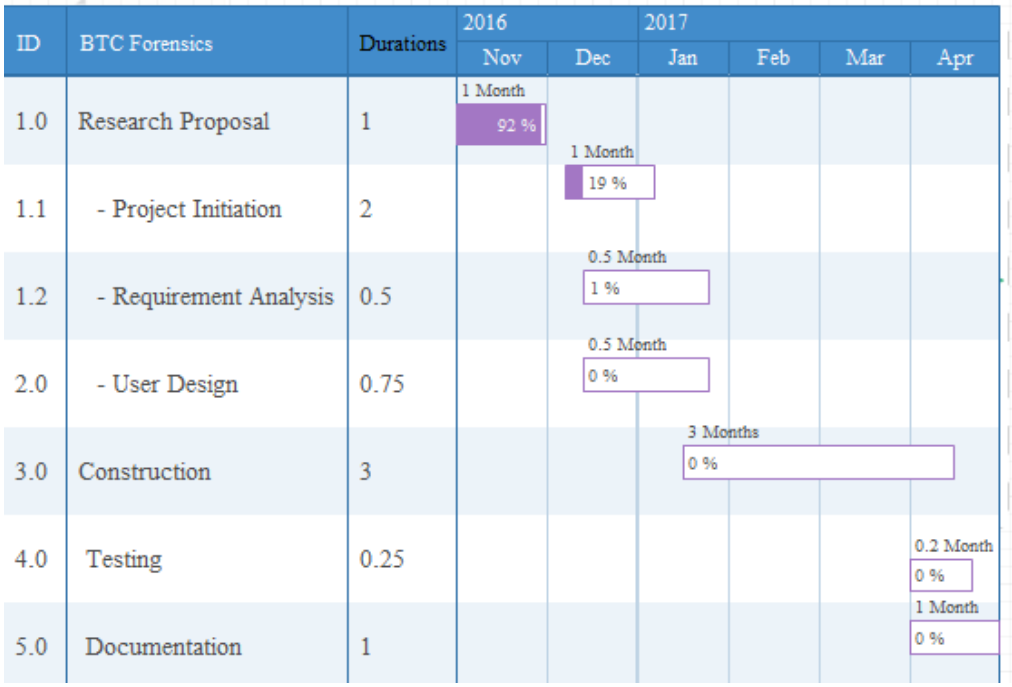
Wiki: Block. (2016, March 4). Retrieved from Wiki: https://en.bitcoin.it/wiki/Block

XAMPP Apache + MariaDB + PHP + Perl. (2016, February 10). Retrieved from Apache Friends: https://www.apachefriends.org/index.html

Xethalis, E. L. (2015). Recent key Bitcoin and virtual currency regulatory and law enforcement developments. *Journal of Investment Compliance, Vol 16, Iss1* , 13-18.

Appendix

Appendix A: Gantt Chart



Appendix B: Sample Output from Blockchain.info

C:\WINDOWS\system32\cmd.exe

```
[2017-04-12 01:31:57,274] [11] [INFO ] - All crawls have completed
Completed crawling all sites
1dice8EMZmqKvrGE4Qc9bUFf9PX3xaYDp
127688421dice97ECuByXAvqXpaYzSaQuP
101561NxaBCFQwejSZbQfWcYNwgqML5wWo
1721LuckyR1fFHEsXYyx5QK4UFzv3PEAep
117421dice9wcMu5hLF4g81u8nioL5mmSH
1dice7fUkz5h4z2wPc1wLMPWgB5mDwKDx
34699071dice7W2AicHosf5EL3GFDUVga7
31435571VayNert3x1KzbpzMGt2qdqrATh
31338961dice6YgEVBf88erBFra9BHf6ZM
1LuckyY9fRzcJre7aou7ZhWVXktxjjBb9S
30261LuckyG4tMMZf64j6ea7JhCz7sDpk6
17211diceDCd27Cc22HV3qPNZKwGnZ8Qwh
19812891dice5wwEZT2u6ESAdUGG6MHgCp
19496511dice7EYzJag7SxkdKXLr8Jn14W
16721841dice1e6pdhLzzWQq7yMidf6j8e
16406471MPxhNkSzeTNTHSZAibMaS8HS1e
16099281Bd5wrFxHYRkk4UCFttcPNMYzqJ
15900121NDpZ2wyFekVezssSXv2tmQgmxc
15891301 8uvwkMJsg9cxFEd1QDFgQpoeXW
15885901 5fXdTyFL1p53qQ8NkrjBqPUbPW
15883811J4yuJFqozxLWTvnExR4Xxe9W4B
15880941 4719bzrTyMvEPcr7ouv9R8utnc
15880251 3h1DP2Boo9TAsenphroACxhNy7
15869251HVpyjYEPwQhvRQ3dL8tGe9kiyd
15868301Bqm5MDo82m1FTxV3qYNUUEKnES
15866801HjDauL2kth6KJUz5vX198Nvp1x
15866151GoK6fv4tZKXFiWL9NuHiwcwsi8
15861891L4EThM6x3Rd2PjNbs1U136FpMq
15854701MSzmVTBaaSpKDARK3VGvP8v7aC
15853061 7NKcZNXqAbxWsTwB1UJHjc9mQG
15851871 3HFqPr9Ceh2aBvcjxNdUycHuFG
15849571 4ChPPM8rPYJeHnw6kMVUDnNNKx
15838541AdN2my8NxvGcisPGYeQTAKdWJu
15838001DpsR91YmHUDTtiuH1pPCuG3RqA
15827031JmcV7G3r8k7ev2EkS84MmsvxGy
15826961PeohaRGaTF8cSzDqP1yYfzDah6
15823261 8czPiA9PcCs7rFTBZnhvNAWuh1
15816071HZHBnH2FbHNWieMxAh4xBPfgfu
15812681 8XSLnBZ8ydMUkaifU6sQBMJzmm
15794451MPerpQzTABa1K2eXQxsQTDSZtD
15794321 2Cf6nCcRtKERh9cQm3Z29c9MWv
15787191 9iVyH1qUxgywY8LJSbpV4VavjZ
14608551dice6DPtUMBpWgv8i4pG8HMjXv
14018241 3vHWR3iLsHeYwT42RnuKYNBoVP
13725151MEe2mebed8wopvy8xyjjHcEQHU
12996831 7BymcHaGRbXGnEzR2m9woUYNf9
12985611GUkazUBpXWdSJ9HbgTapAH7uyb
3Cs6K12978581KYXrw4FtkmomfS4iyVXUS
12975841Po4J4SNyJuGnMGYJfGTXLEvGgA
12971651 4wXrm49HxggbdQ6RGfWY8qghGE
12957601KumbRsTcA6UNqU2MHEfqEFnAZY
3w3izR12956391Lsqcv4cg5zUctNi2qwNx
1GeBboSUJ12955301 2ucu9bHLe2w2ahjib
3xmdcGkgUmX4A12953521Bet32kBtZzXVi
11741431dicec9k7KpmQaA8Uc8aCCxfWnw
11139121Fi57hAqyYYwaQVdA7a9qSKfiuk
31G310483141diceeEArYHebwOZhyr5G9A
```

Appendix C: Sites.txt

Sites.txt

```
1  https://blockchain.info/popular-addresses
2  http://Coin.mx
3  http://Crypto-Trade.com
4  http://VaultOfSatoshi.com
5  http://localbitcoins.com
6  http://huobi.com
7  http://localhost:8080/myaddress/mshome.html/
```